Using LLMs to Build a Database of Climate Extreme Impacts

Ni Li[*1] **Shorouq Zahra**[†¶1] **Mariana Madruga de Brito**[‡] **Clare Marie Flynn**[§¶]
Olof Görnerup[†¶] **Koffi Worou**[§¶] **Murathan Kurfalı**[†¶] **Chanjuan Meng**[†¶]
Wim Thiery[*] **Jakob Zscheischler**[‡] **Gabriele Messori**[§¶] **Joakim Nivre**[†§¶]

[*]Vrije Universiteit Brussel [†]RISE Research Institutes of Sweden
[‡]Helmholtz Centre for Environmental Research – UFZ [§]Uppsala University
[¶]Swedish Centre for Impacts of Climate Extremes (climes)

Abstract

To better understand how extreme climate events impact society, we need to increase the availability of accurate and comprehensive information about these impacts. We propose a method for building large-scale databases of climate extreme impacts from online textual sources, using LLMs for information extraction in combination with more traditional NLP techniques to improve accuracy and consistency. We evaluate the method against a small benchmark database created by human experts and find that extraction accuracy varies for different types of information. We compare three different LLMs and find that, while the commercial GPT-4 model gives the best performance overall, the open-source models Mistral and Mixtral are competitive for some types of information.

1 Introduction

Increasingly frequent and intense extreme climate events pose significant threats globally at both individual and collective levels. However, we still do not have a robust understanding of how extreme climate events impact society, which in turn hinders impact forecasting, early warning, and disaster risk management (de Brito et al., 2024). Accurate impact information is crucial for identifying areas disproportionately affected (Hammond et al., 2015), enabling targeted allocation of climate adaptation efforts. Such data can also provide support for the evaluation of whether adaptation measures effectively reduce loss and damage from climate extremes (Kreibich et al., 2023).

Existing publicly accessible global climate impact databases suffer from incomplete, inconsistent and/or biased data (Tschumi and Zscheischler, 2020; Panwar and Sen, 2020). One of the most used natural hazards-related impact databases is EM-DAT (Delforge et al., 2023).[2] While EM-DAT is an extremely valuable database, events are often assigned non-standardized spatial information: from city to country scales, or geophysical areas without clear formal boundaries. Similarly, temporal specifications may be a date range in days, in months or only a year. The impacts from a single physical event may further be listed under multiple separate entries if affecting an extended area. Moreover, events in both developed and developing countries are likely underreported (Harrington and Otto, 2020). Many climate extremes also lack impact information in one or multiple categories (Jones et al., 2022). Some of these constraints are also shared by other multi-hazard, multi-impact databases, such as DesInventar (UNISDR, n.d.). Single-hazards databases (e.g., Paprotny et al., 2023) and/or databases focusing on national spatial scales (Sodoge et al., 2023) have better coverage and completeness, yet they typically cannot be easily updated or scaled to multiple hazards or regions. Moreover, they adopt differing impact categories and event definitions, preventing any multi-hazard impact analyses.

In this paper, we propose a method for constructing a database of climate extreme impacts from online textual sources, using natural language processing (NLP). This has the potential to address the above-mentioned database limitations, ensuring broad spatiotemporal coverage, standardisation of information and easy updating. Our approach leverages the power of large language models (LLMs) and in-context learning to extract semi-structured information, which is normalized and refined in post-processing and stored in a relational database. A crucial step in the refinement process is geoparsing, which maps place names to geographical entities in order to enhance the usefulness of the database for researchers. Another important feature of the database is that we store the actual text from which the information has been extracted, allowing users to trace sources and validate the information.

[1]Equal contribution of first two authors.
[2]https://www.emdat.be

An empirical evaluation based on a benchmark database created by human experts for 170 extreme climate events shows that extraction accuracy varies for different types of information. While the main event category (such as "Flood" or "Wildfire") and the number of people killed can usually be determined with high accuracy, geographic locations and total economic damage are harder to extract reliably. A comparison of three different LLMs shows that the commercial GPT-4 model gives superior performance overall, but the open-source models Mistral and Mixtral give competitive results for some information categories.

2 Database Design

The first step towards an information extraction system for climate extreme impacts is the design of a database schema, which defines what type of information should be extracted and how this information should be formally represented. An important consideration here is compatibility with existing de facto standards in the field, and we have therefore chosen to base our categories mainly on the existing EM-DAT database (Delforge et al., 2023), while trying to overcome some of its limitations.

Figure 1 gives a schematic overview of the kind of schema used in our system. The fundamental entity is an *event*, which is a climate-related extreme such as a storm or a heatwave. Each such event must be associated with information about its *location*, *time* and *event category*. This is a basic requirement, because information about impacts that cannot be located in space and time is of no use to scientists, but in order for an event to be included in the database, there must also be some information about its impacts.

By *impacts* we understand the socio-economic impacts of climate extremes, that is, the negative repercussions of such events on society (de Brito et al., 2024). As shown in Figure 1, we subdivide these into (a) direct impacts to persons, such as the number of fatalities and of persons being injured, displaced or homeless, and (b) material and economic damage, such as insured and total economic damage, and building damage. The specific impact types are chosen to ensure compatibility with existing impact databases, in particular, EM-DAT:

- Deaths: Number of people killed.
- Injured: Number of people injured.
- Displaced: Number of people displaced.
- Homeless: Number of people made homeless.
- Affected: Number of people affected.
- Insured damage: Cost of insured damage.
- Total damage: Cost of total damage.
- Buildings: Number of buildings damaged.

Since an event may have different impacts at different times and locations, the value for each impact type is a set of triples $\langle val, loc, time \rangle$, where val is a numerical value (number or cost, depending on the type), while loc and $time$ are specifications of a location and a time. In addition, we provide a global numerical value for the event as a whole. Finally, to allow users to trace the information source, we store both a global document reference and specific text passages for each extracted information item. Below, we describe in more detail how information about location, time, event category, number, and cost is represented in the database.

Location A location is specified across multiple fields encoding different levels of information and as standardized as possible. These fields are:

- `Name` (string): This field contains a standardized name of the location. This can be the international name, the official English name, or the Wikipedia article title of that location, whichever is available on OpenStreetMap (OpenStreetMap contributors, 2017a) and in that order of preference.
- `Type` (string): This field represents the type of the location as listed on OpenStreetMap, which essentially follows the ISO 14819-3 standard (OpenStreetMap contributors, 2017b). Countries would often be of type *administrative*.
- `GeoJSON` (JSON object): GeoJSON is a format for encoding geographic data structures that is based on JSON (JavaScript Object Notation). Each location is represented by one of these planar geometric features: `Point`, `LineString`, `Polygon`, `MultiPoint`, `MultiLineString`, or `MultiPolygon`. Countries are usually represented by the geometry type `MultiPolygon` whereas straits or rivers may be represented as type `LineString`. These geometric shapes are pulled directly from OpenStreetMap and enable users to visualize impact locations on a world map.
- `GID` (unique identifier): GID is a unique ID used by the Database of Global Administrative Areas (GADM) (Global Administrative Areas, 2012) to represent countries and their administrative areas.

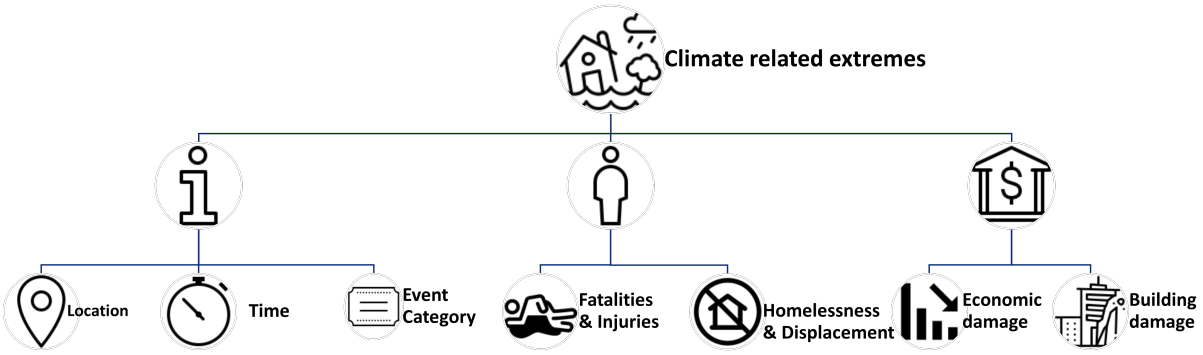

Figure 1: Simplified schema for a database of climate extreme impacts.

Since an event or a reported impact may affect multiple locations, each of the fields above in fact contains a set of values (`Name`, `Type`, `GeoJSON`, and `GID`, respectively) for each location.

Time The time of an event is specified by a start date and an end date, which are the same if the event took place within a single day. (We do not consider shorter time periods than one day.) The dates are specified in `YYYY-MM-DD` format, where the year is strictly required, while the month and day fields are nullable in case the information is not available. Formally, this is represented by a tuple $time = \langle syear, smon, sday, eyear, emon, eday \rangle$, where $syear$ and $eyear$ are 4-digit integers, while $smon$, $sday$, $emon$, and $eday$ are 2-digit integers or `NULL`.

Event Category The event category is specified by a string value from the following closed set:

- `Drought`
- `Extreme Temperature`
- `Flood`
- `Wildfire`
- `Tornado`
- `Tropical Storm/Cyclone`
- `Extratropical Storm/Cyclone`

The selection of categories has been made with compatibility with existing resources in mind. Flood is a separate event category, but can also result from a tropical or extratropical cyclone. The reasoning for also having it as a separate category is that floods can be caused by a variety of other factors, from convective summer rain to rapid snowmelt.

Number Several impact types[3] are specified by giving the number of people (or buildings) affected in some way. Such numbers can be reported in textual sources in a variety of ways, including an exact number (e.g., "23"), a closed or open interval range (e.g., "20–25", "over 100"), or some other approximation (e.g., "around 100", "hundreds"). To facilitate automatic processing of the information, we want to avoid string representations, which have to be parsed to be interpreted, and therefore standardize the different values to a uniform representation $num = \langle min, max, app \rangle$, where min and max are the minimum and maximum of a value range, and app is a boolean value indicating whether the information is approximate or uncertain. This representation allows us to capture the most commonly occurring specifications as follows:

- Exact numbers like "25" are mapped to a range with $min = max$: $\langle 25, 25, \texttt{False} \rangle$
- Exact ranges like "20–25" are mapped to a range with $min \neq max$: $\langle 20, 25, \texttt{False} \rangle$
- Open ranges and approximations are mapped to suitable ranges with $app = \text{True}$. Thus, "around 100" is mapped to $\langle 100, 100, \texttt{True} \rangle$, "hundreds" is mapped to $\langle 200, 900, \texttt{True} \rangle$, and "over 100" is mapped to $\langle 100, 199, \texttt{True} \rangle$.

Cost Insured damage and total damage are specified as a monetary cost, that is, as a specific amount in a specific currency, for example, "2,500,000 USD". Formally, this is represented in the database by a triple $cost = \langle min, max, currency \rangle$, where min and max are the minimum and maximum of a value range (as for Number above), and $currency$ is an ISO 4217 currency code.

3 Information Extraction

Our method for populating a database of climate extreme impacts based on information extraction from online textual sources uses a pipeline consisting of three main components, as illustrated in Figure 2. The first component performs document

[3] Deaths, Injured, Displaced, Homeless, Affected, Buildings.

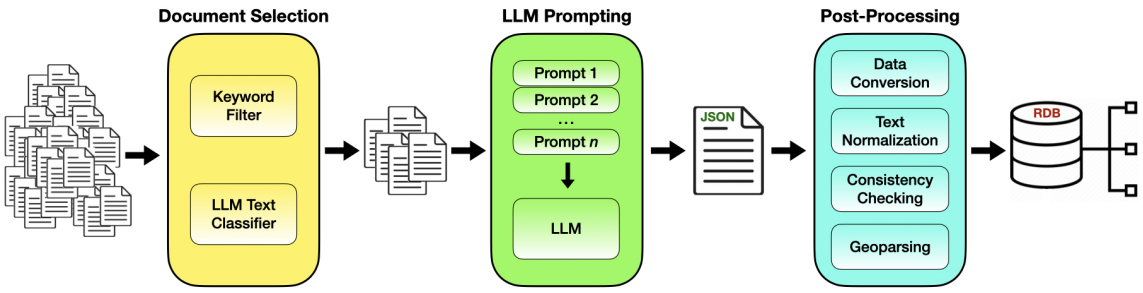

Figure 2: Pipeline with three main modules: document selection, LLM prompting, and post-processing.

selection using web scraping with keyword filtering and an LLM-based text classifier. The second component uses LLM prompting to extract information about extreme climate events and their impacts, storing the result in a semi-structured format. The third component post-processes the semi-structured information by converting all information items to their correct data type, normalizing all text elements, performing various consistency checks, and mapping location names to geographical entities, before storing the result in a relational database. Below we describe each of the three components in more depth.

3.1 Document Selection

Information about the impacts of climate extremes can be found in diverse sources on the internet, and our system is capable of handling arbitrary text documents, although we have initially targeted articles from English Wikipedia. To select relevant articles, we use a two-step approach, where the first step uses a simple keyword filter and the second step uses a domain-specific text classifier.

The list of keywords used in the first step was hand-crafted by domain experts in our team with the goal of covering all major event categories in the database. The full list of keywords can be found in Appendix A. The text classifier used in the second step was created by fine-tuning the pre-trained English BERT model (Devlin et al., 2019) on a small corpus of 300 Wikipedia articles, containing 248 relevant and 52 irrelevant articles.[4] Using 150 articles for training, 100 articles for development and 50 articles for testing, we obtained an F_1-score of 98.8 on the test set (precision 97.7, recall 100.0).

We applied the document selection to all of English Wikipedia, where the first step resulted in a selection of 30,085 articles, of which 4,900 were classified as relevant in the second step. One of the authors then manually went through all 30,085 articles, checking only the first sentence of each article, and in this way identified 184 false positives in the selection of 4,900 articles and another 330 false negatives in the remaining 25,185 articles. Discounting the 300 articles used to train the classifier, this corresponds to an F_1-score of 94.5 (precision 96.0, recall 93.0). Although this is not a rigorous evaluation of the method, and it is not clear how well the method would work for other types of documents than Wikipedia articles, the results nevertheless strongly indicate that it is a feasible task to identify relevant documents for further processing.

3.2 LLM Prompting

In the core component of our information extraction pipeline, we feed articles to an LLM together with a sequence of prompts designed to extract information corresponding to the different fields of our database. For basic information about the event, such as location, time, and event category, we pose two questions, one for the required piece of information and one for the text passage where this information can be found (to be stored in the database for traceability and validation). For the different impact types, we use more complex prompts to extract information at the global event level as well as for specific times and locations if available. To facilitate post-processing, we instruct the LLM to provide output in JSON. A selection of representative prompts can be found in Appendix B.

During the development and prompt engineering process, we have so far relied exclusively on GPT-4 (OpenAI et al., 2024) as the LLM, but our experimental evaluation includes a comparison with two popular open-source models: Mistral (Jiang et al., 2023) and Mixtral (Jiang et al., 2024).

[4]For articles longer than 512 tokens, only the first 512 tokens were used.

3.3 Post-Processing

Although the JSON output produced by LLMs tends to be well-formed as regards the global structure, the detailed information about event properties and impacts is often inconsistently formatted and sometimes of the wrong data type. It is, therefore, necessary to perform various types of post-processing to ensure that the input to the database is well-typed and consistently formatted. For location information, the post-processing involves not only the normalization of geographical names but also mapping these names to types, GeoJSON objects, and unique GADM IDs (called GIDs) for various levels of subdivisions (Global Administrative Areas, 2012). Below we describe the most important post-processing steps in more detail.

Location The LLMs are prompted to produce a list of both countries and smaller, more fine-grained areas within a country (if mentioned) for each event. The extracted areas sometimes appear in an alternative spelling or describe broader regions by their local or colloquial names rather than by their official administrative titles.

Several steps are taken to normalize locations. In general, locations are disambiguated and normalized using OpenStreetMap (OpenStreetMap contributors, 2017a) or using the UNSD dataset[5] for mapping geographical regions (such as "North America") to a list of countries. When querying OpenStreetMap, we limit the search for a location within a certain country (if present) which greatly improves the normalization results. Administrative or natural areas (such as cities, national parks, or islands) are preferred, while undesirable location "types" (OpenStreetMap contributors, 2017b) (such as clinics or hospitals or car parks) are ignored. Results are sorted in ascending order by their search rank (Nominatim contributors, 2014) and the topmost result is returned. From OpenStreetMap, a standardized international name and GeoJSON object are retrieved for each location.

On top of normalizing with OpenStreetMap, we also match locations with a unique ID called GID from GADM (Global Administrative Areas, 2012) for all available levels (where level 0 is the "country" level, and each level further up divides a single country into smaller administrative subdivisions).

Time The LLM extraction outputs dates in a variety of formats or locales. Since extreme climate events may span several months or even years, these extracted dates may appear without a day or month. Some examples of a variety of date formats that are extracted by the LLM: "21 January 2008","2018-07-17", "1996", and "March 2015". These are normalized using a data parsing library in Python (`dateparser` (DateParser contributors, 2024)).

Number/Cost We find that the LLM extraction output (whether the total number of people or the total amount of monetary damage) is sometimes in the form of a phrase, such as "None reported", "At least 1,152", or "EUR54 billion", rather than a single number or range: "0", "1152", or "54000000000", respectively. If the LLMs output a single number, this is extracted and parsed with the correct locale to account for the decimal separators (such as a comma or period, which differs by country). LLM outputs that mix numbers with words are first cleaned of currency symbols. Digits and spelled-out numbers are then normalized with the help of Python libraries that convert natural language texts to numbers[6] and vice versa.[7] Finally, they are parsed using a rule-based approach that considers the part-of-speech tags and entities predicted by SpaCy's English transformer pipeline model,[8] as well as the presence of scales (such as "million", "thousand"; but also other scales like "crore", or "lakh" from the Indian numbering system, which appears in the development set).

If two numbers appear in the text, we assume that they represent a range and extract them with a similar rule-based approach based on the part-of-speech tags and entities (from SpaCy). In addition, we use a rule-based approach to infer whether or not the given range of numbers is an estimate or an exact number. Finally, we employ a small list of phrases that directly map to a numeric output: "None" translates to $\langle 0, 0, \texttt{False} \rangle$ (where `False` means the number is exact) while "tens of casualties" is mapped to $\langle 20, 90, \texttt{True} \rangle$ (where `True` means the number is an approximation).

4 Evaluation

We evaluate our method using development and test data annotated by domain experts in our team. The experimental evaluation involves a comparison of three different models in the second step of

[5] https://unstats.un.org/UNSDWebsite/

[6] https://github.com/allo-media/text2num

[7] https://github.com/savoirfairelinux/num2words

[8] https://spacy.io/models/en#en_core_web_trf

Source	Articles	Single	Multi
Artemis	57	46 (81%)	11 (19%)
Wikipedia	243	240 (99%)	3 (1%)
Total	300	286 (95%)	14 (5%)

Table 1: Overview statistics of the articles used for the benchmark database, including media source type and breakdown of single- vs. multi-event articles.

the pipeline, while keeping the input data and post-processing constant. The three models are GPT-4[9] (OpenAI et al., 2024), Mistral[10] (Jiang et al., 2023), and Mixtral[11] (Jiang et al., 2024). The same prompts are used for all models (cf. Appendix B), except for an additional final sentence to ensure responses are strictly in JSON format for the Mistral models, to overcome their tendency to produce additional comments. Below, we first describe the data annotation and define the evaluation metrics before reporting and discussing our experimental results.

4.1 Data Annotation

Our annotated data is based on documents in English taken from Wikipedia and Artemis.[12] Artemis is a media service of the insurance industry and focuses on catastrophe bonds, insurance-linked securities, reinsurance, and risk transfer, while regular Wikipedia articles were used. The Artemis and Wikipedia texts were obtained through web scraping based on a keyword filter (cf. Section 3.1), such that both relevant and irrelevant documents were included. However, for the purpose of this article, where we do not evaluate the document selection step, only relevant documents have been included.

The annotation was performed in two steps. First, spans in the actual text were labeled with categories corresponding to event categories, times, locations, and all the impact types defined in the database schema (cf. Section 2). Secondly, for each extreme climate event described in an article, a database record was created. In the evaluation reported below, we only make use of the output of the second step, which we refer to as the benchmark database.

The benchmark database is based on 300 unique articles, statistics of which are shown in Table 1. This includes 57 unique articles from Artemis and 243 from Wikipedia, representing 19% and 81% of the unique articles, respectively. These articles can be further classified as single- or multi-events. A single-event article describes only one extreme climate event, whereas a multi-event article reports on several such events. The Wikipedia source article *2021 European Floods*[13] exemplifies a single-event article for the floods that devastated much of Europe in the summer of 2021. While the floods were extensive and affected multiple countries over a prolonged period of time, they were associated with a single main climatic driver in the form of heavy precipitation from a weather system, and are thus physically a single extreme event. The Artemis article *Storm Eberhard industry loss estimated up to EUR 1.5bn by AIR*[14] demonstrates a multi-event article covering the European winter windstorms Dragi-Eberhard and Freya (Bennet). Most unique articles are classified as single-event (286 articles or 95%), rather than multi-event (14 articles or 5%). More Artemis-sourced articles are classified as multi-event relative to Wikipedia-sourced (19% and 1%, respectively), but a clear majority of articles from both sources are single-event.

The benchmark database contains, in total, 289 events, defined as an extreme climate event belonging to one of our seven event categories, occurring at a specified date or date range and geographic location, typically at the country level. The main event for the 2021 European Floods, for example, is defined as a flood event type, affecting the countries the United Kingdom, Austria, Belgium, Croatia, Germany, Italy, Luxembourg, the Netherlands, Switzerland, and Romania, and over the date range 2021-07-12 to 2021-07-25. 199 events, or 61%, only have impacts specified for the event as a whole, while 90 (31%) have impact specifications for specific times or locations. For example, flood impact information for a specific country within the country list of the 2021 European Floods, or a specific location within a single country from this list, is specified separately. In the first evaluation, we only include impacts at the main event level.

The benchmark database covers a long time record: 1287-12-13 to 2023-02-17, though the majority of events occur in the 20th and 21st centuries.

[9] GPT-4-turbo-2024-04-09; GPT-3.5-turbo-1106 for articles with a length shorter than 32,500 characters, and for time information.

[10] mistralai/Mistral-7B-Instruct-v0.2

[11] mistralai/Mixtral-8x7B-Instruct-v0.1

[12] https://www.artemis.bm

[13] https://en.wikipedia.org/wiki?curid=68241636

[14] https://www.artemis.bm/news/storm-eberhard-industry-loss-estimated-up-to-eur-1-5bn-by-air/

The nine events that do not occur during or after the year 1900 include the 1287 St. Lucia's Flood event and eight events in the late 18th and late 19th centuries. Further, 92% of the events occur during or after the year 1960. Considering geographical regions, most events occurred in North America, followed by Asia and Europe, while the fewest were found in South America. Among event categories, tropical storms are by far the most frequent, followed by floods and extratropical storms, while extreme temperatures, drought, wildfires and tornados are less frequent. Droughts are a difficult event category for our database schema, as their impacts are often not specified using concepts defined in the database. More information about the distribution over geographical regions and event categories can be found in Appendix C.

For the experimental evaluation reported below, we use 100 events as development data and 170 events as test data. The proportion of Wikipedia articles is 84% (84/100) in the development set and 93% (158/170) in the test set.

4.2 Evaluation Metrics

The information extracted for each extreme climate event is quite complex, and evaluation is therefore not completely straightforward. To obtain an aggregated score for each event, as well as scores for specific fields, we define a difference metric for each field, ranging from 0 to 1 (where lower is better), and derive an aggregated score as a weighted sum of field-specific scores:

$$D(a,r) := \frac{1}{n}\sum_i w_i d_i(a_i, r_i) \qquad (1)$$

$D(a,r)$ is the difference between an annotated (benchmark) record a and a retrieved record r, with weights w_i and difference metrics d_i of fields i, where n is the number of fields. In this way, the relative influence of each field can be adjusted using its weight if we regard some fields as more important. For the evaluation in this paper, however, we use uniform weights for all fields.

The difference metrics for specific fields are defined in terms of metrics for the following basic types: numbers, strings, booleans, and sets, each in the range $[0,1]$:

- For (non-negative) numbers:

$$d_n(a,r) := \begin{cases} 0, & \text{if } a = r \\ \frac{|a-r|}{a+r}, & \text{otherwise} \end{cases} \qquad (2)$$

- For strings and booleans:

$$d_{t,b}(a,r) := \begin{cases} 0, & \text{if } a = r \\ 1, & \text{otherwise} \end{cases} \qquad (3)$$

- For sets:

$$d_s(a,r) := 1 - \frac{|a \cap r|}{|a \cup r|} \qquad (4)$$

The use of these tailored metrics, rather than standard accuracy, recall, or precision metrics, is motivated by the database's intended use in modeling climate extremes and their impacts. For example, if the correct number of deaths is 10, then a prediction of 11 is an almost negligible error, while a prediction of 100 is severe. With the current metric, these predictions get a normalized difference score of 0.048 and 0.818, respectively.

Our evaluation in this paper is limited to five representative database fields, for which the difference metrics are defined as follows (cf. Section 2):

- **Location:** A set of normalized country names, evaluated using the set metric $d_s(a,r)$.
- **Time:** A sextuple of numbers, representing the start and end date, each evaluated using the number metric $d_n(a,r)$.
- **Event Category:** A category label, evaluated using the string metric $d_{t,b}(a,r)$.
- **Deaths:** Two (possibly) identical numbers, representing the minimum and maximum value of a range, each evaluated using the number metric $d_n(a,r)$.[15]
- **Total Damage:** A triple of values, representing the minimum and maximum value of the amount, and the currency, evaluated using the number metric $d_n(a,r)$ (min, max) and string metric $d_{t,b}(a,r)$ (currency).

Although this is a limited subset of the database fields, it nevertheless includes all major types of fields, including one person-oriented and one cost-oriented impact.

4.3 Experimental Results

Table 2 presents the performance of three language models across the selected database fields. The average scores indicate that GPT-4 consistently outperforms the other models with robust performance across both Wikipedia and Artemis articles. The

[15] Note that we do not evaluate the boolean value indicating whether the numerical values are approximate.

Category	GPT-4			Mistral (7B)			Mixtral (8x7B)		
	Tot	**Wik**	**Art**	**Tot**	**Wik**	**Art**	**Tot**	**Wik**	**Art**
Event Category	0.106	0.108	0.083	0.088	0.089	0.083	0.100	0.101	0.083
Location	0.295	0.302	0.216	0.452	0.438	0.647	0.446	0.440	0.526
Start-Year	0.041	0.044	0.000	0.753	0.804	0.083	0.141	0.139	0.167
Start-Month	0.043	0.046	0.000	0.753	0.804	0.083	0.150	0.149	0.167
Start-Day	0.047	0.051	0.000	0.762	0.813	0.093	0.167	0.167	0.168
End-Year	0.024	0.025	0.000	0.771	0.810	0.250	0.189	0.184	0.250
End-Month	0.027	0.028	0.012	0.772	0.811	0.262	0.196	0.191	0.261
End-Day	0.039	0.042	0.004	0.776	0.817	0.250	0.227	0.225	0.250
Deaths-Min	0.046	0.036	0.167	0.199	0.202	0.167	0.188	0.189	0.167
Deaths-Max	0.046	0.037	0.167	0.209	0.212	0.167	0.183	0.185	0.167
Damage-Min	0.151	0.099	0.833	0.611	0.626	0.417	0.454	0.463	0.333
Damage-Max	0.151	0.099	0.833	0.600	0.614	0.417	0.454	0.463	0.333
Damage-Currency	0.129	0.076	0.833	0.294	0.241	1.000	0.394	0.367	0.750
Total Event	0.082	0.071	0.225	0.503	0.520	0.280	0.235	0.233	0.259

Table 2: Results on the test set with three different LLMs: GPT-4, Mixtral, Mistral. Average difference over all events (Tot) and separately for Wikipedia (Wik) and Artemis (Art) articles. For start and end dates, we evaluate year, month and day separately; similarly for minimum and maximum values for deaths and total damage, and currency for total damage. The total event score is the unweighted mean of all the individual field scores.

only noticeable discrepancy is in the damage category, where GPT-4's performance drops significantly in the Artemis articles. Notably, we find that LLMs tend to confuse insured damage with total damage in Artemis articles, whereas Wikipedia articles often present the total economic damage clearly in the information box, which explains the large divergence in error rates between Artemis and Wikipedia articles. For most other fields, the error rate for GPT-4 is around or below 0.1. The only exception is Location, where scores are in the 0.2–0.3 range.

In contrast to GPT-4, Mistral exhibits significantly higher error rates and more variation across Wikipedia and Artemis. It especially struggles with extracting dates and damages, with error rates between 0.6 and 0.77.[16] Interestingly, it achieves much better performance on Artemis, where the error rate is almost half of that for Wikipedia. Mixtral is found to be a better alternative to GPT-4 with consistent performance, although not as accurate. It performs significantly better than Mistral in the date categories while still struggling with damage. Unlike Mistral, Mixtral's performance is more stable across Wikipedia and Artemis. However, it is interesting that, unlike GPT-4, the Mistral models perform better or similarly on Artemis, suggesting a potential overfit of the prompts for GPT-4 and Wikipedia. All models have similar performance on development and test sets, which suggests that there is no overfitting for prompts in general.[17]

One of the reasons behind the lower performance of open-source LLMs is their inability to output valid JSON files, which inevitably leads to data loss. In the test set of 170 events, we asked the models to generate 850 JSON files (170 events multiplied by 5 prompts each), and approximately 20% of these were not valid JSON files. We managed to recover half of these invalid JSON files through post-processing in the case of Mistral, and around 65% in the case of Mixtral. However, this does not imply that the remaining invalid JSON files are without value; they still store meaningful information, but it is not possible to extract this data due to the formatting issues.

In terms of specific fields, the event category is the easiest one to identify, with all models achieving scores around 0.1 (and with the Mistral model interestingly outperforming the two other models), whereas location and damage-related fields are the most challenging. The error rate for location is about 0.2–0.5 across the models and article types, and an error analysis reveals that several errors

[16] For dates, this is mainly due to erratic or invalid JSON formatting in the LLM output, which leads to data loss or incorrect normalization.

[17] Development set results can be found in Appendix D.

are caused by locations that cover multiple countries, in particular archiepelagoes like the Caroline Islands and the Mariana Islands, which are not retrieved correctly by the LLMs. For the total damage field, a challenge is that this is often reported by less exact phrases, such as "minimal", ">$1.8 million", compared to other fields. Increasing the accuracy of these fields is likely to require a combination of more advanced prompting strategies and improved post-processing.

5 Related Work

The notion of using NLP for extracting impact information from textual data is rapidly gaining traction in the fields of climate and impact science. While no previous work has attempted to build a global multi-hazard database, such as the one that we are presenting here, there have been a number of implementations of NLP approaches in more targeted contexts. For instance, de Brito et al. (2020) extract and classify impact statements in newspaper articles for the 2018/19 German drought. This line of work is continued by Sodoge et al. (2023) and Alencar et al. (2024), who use supervised classification models to extract information from newspaper articles on the different socio-economic impacts of droughts in Germany. NLP approaches have also been applied to social media, for example by Zhang et al. (2021), who use a BERT model to identify mentions of seven different types of drought impacts in Twitter data originating in California, United States. Other authors have used automated processing of textual data to provide a broader categorisation of climate extremes going beyond categorical impacts, notably Kahle et al. (2022), who map the course, consequences, and aftermaths of the 2021 European floods. Finally, as a direct precursor of the information extraction approach presented in this paper, we mention Li (2023), who focuses on Wikipedia articles and URLs to extract impacts of multiple classes of climate extremes, achieving 86% accuracy for time and 92% for location with GPT-3.5, surpassing the performance of a BERT model.

6 Discussion and Conclusion

We have presented the first evaluation of an LLM-based system for building a database of climate extreme impacts. The results show that this is a challenging task, especially for certain types of information, and that LLMs still need to be supported by more traditional NLP techniques to ensure correct data typing and consistency. Our comparison of different LLMs indicates that open-source models match the performance of GPT-4 on specific information types (in particular the main event category), and it is likely that the results can be improved further through model-specific prompt engineering and better pre- and post-processing.

Even discounting inaccuracies introduced by the LLMs, the quality of the database depends on the correctness of the data presented in the Wikipedia and Artemis articles. The issue of potentially incorrect or incomplete impact data is shared with other current state-of-the-art global impact datasets (e.g. DesInventar and EM-DAT; Panwar and Sen, 2020; Jones et al., 2022). In this respect, it is crucial to underscore that there is often no ground truth for impacts of a specific event, as many impacts cannot be or are not directly measured, but rather are estimated.

Despite the inherent biases in using Wikipedia and Artemis as data sources, our approach presents several advances upon existing global impact datasets that are routinely used. Existing datasets typically include manual and unsystematic compilation steps, and do not connect entries to specific sources, thus hindering validation. In contrast, our proposed database enables users to trace each entry back to a specific textual source. Moreover, unlike most current impact databases we include ranges where no precise numbers are reported in our sources or where multiple estimates are quoted, thus facilitating uncertainty quantification. Finally, the highly automated pipeline that we developed enables frequent updates of the database, for example, if new impact information or data sources become available.

We nonetheless recognize that several additional steps may further facilitate the use of our database in research, notably connecting entries to observed environmental variables (e.g. water levels, wind speeds, temperatures). We thus conclude that, notwithstanding practical and technical challenges, LLMs are a promising tool to develop a new generation of databases of climate extreme impacts.

Limitations

The study presented in this paper has a number of limitations that should be considered when interpreting its results. The evaluation only covers a limited number of fields in the database schema and

is based on a relatively small test set due to a lack of resources. The test set is furthermore skewed in several respects, in particular concerning article types, event categories and geographical locations. Moreover, the comparison of LLMs is likely to be biased by the fact that prompts were engineered for GPT-4 and then applied with minimal adaptation to Mixtral and Mistral. Finally, the fact that only documents in English are considered constitutes a further limitation. The evaluation results must, therefore, be interpreted with caution, and further studies are needed to assess to what extent they can be generalized to other settings, models, languages, and data distributions.

Ethics Statement

We do not foresee this paper raising any major ethical issues. It only uses public data sets with no personal or otherwise sensitive information, and all annotation has been performed by team members and students who have been compensated fairly for their efforts. Nonetheless, due to a combination of factors including the use of data in English only, the selection of extreme events is biased towards certain geographical regions. The extension of this work to other languages is therefore important to mitigate this bias.

Acknowledgments

The research presented in this paper was supported by the Swedish Research Council (grants no. 2022-02909, 2022-03448 and 2022-06599). Ni Li is supported by the VUB Research Council in the framework of a EUTOPIA inter-university co-tutelle PhD program between the Vrije Universiteit Brussel, Belgium, and the Technische Universität Dresden, Germany. The EUTOPIA alliance is part of the European Universities Initiatives co-funded by the European Union. We thank two anonymous reviewers for their constructive comments.

References

Pedro H L Alencar, Jan Sodoge, Eva Paton, and Mariana Madruga de Brito. 2024. Flash droughts and their impacts – using newspaper articles to assess the perceived consequences of rapidly emerging droughts. *Environmental Research Letters*.

DateParser contributors. 2024. Dateparser – python parser for human readable dates. `https://github.com/scrapinghub/dateparser/tree/master`.

Mariana Madruga de Brito, Christian Kuhlicke, and Andreas Marx. 2020. Near-real-time drought impact assessment: A text mining approach on the 2018/19 drought in Germany. *Environmental Research Letters*, 15(10):1040a9.

Mariana Madruga de Brito, Jan Sodoge, Alexander Fekete, Michael Hagenlocher, Elko Koks, Christian Kuhlicke, Gabriele Messori, Marleen de Ruiter, Pia-Johanna Schweizer, and Philip J. Ward. 2024. Uncovering the dynamics of multi-sector impacts of hydrological extremes: A methods overview. *Earth's Future*, 12(1):e2023EF003906.

Damien Delforge, Valentin Wathelet, Regina Below, Cinzia Lanfredi Sofial, Margo Tonneliere, Joris van Loenhout, and Niko Speybroeck. 2023. EM-DAT: The emergency events database. 10.21203/rs.3.rs-3807553/v1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Global Administrative Areas. 2012. GADM database of Global Administrative Areas, version 2.0. [online. URL: www.gadm.org.

Michael J Hammond, Albert S Chen, Slobodan Djordjević, David Butler, and Ole Mark. 2015. Urban flood impact assessment: A state-of-the-art review. *Urban Water Journal*, 12(1):14–29.

Luke J. Harrington and Friederike. E. L. Otto. 2020. Reconciling theory with the reality of African heatwaves. *Nature Climate Change*, 10(9):796–798.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antonial, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Rebecca Louise Jones, Debarati Guha-Sapir, and Sandy Tubeuf. 2022. Human and economic impacts of natural disasters: Can we trust the global data? *Scientific data*, 9(1):572.

Michael Kahle, Michael Kempf, Brice Martin, and Rüdiger Glaser. 2022. Classifying the 2021 'ahrtal'flood event using hermeneutic interpretation, natural language processing, and instrumental data analyses. *Environmental Research Communications*, 4(5):051002.

Heidi Kreibich, Kai Schröter, Giuliano Di Baldassarre, Anne F. Van Loon, Maurizio Mazzoleni, G.uta W. Abeshu, Svetlana Agafonova, Amir AghaKouchak, Hafzullah Aksoy, Camila Alvarez-Garreton, Blanca Aznar, Laila Balkhi, Marlies H. Barendrecht, Sylvain Biancamaria, Liduin Bos-Burgering, Chris Bradley, Yus Budiyono, Wouter Buytaert, Lucinda Capewell, Hayley Carlson, Yonca Cavus, Anaïs Couasnon, Gemma Coxon, Ioannis Daliakopoulos, Marleen C. de Ruiter, Clare Delus, Mathilde Erfurt, Giuseppe Esposito, Didier François, Frédéric Frappart, Jim Freer, Natalia Frolova, Animesh K. Gain, Manolis Grillakis, Jordi O. Grima, Diego A. Guzmán, Laurie S. Huning, Monica Ionita, Maxim Kharlamov, Dao N. Khoi, Natalie Kieboom, Maria Kireeva, Aristeidis Koutroulis, Waldo Lavado-Casimiro, Hong-Yi Li, Maria C. LLasat, David Macdonald, Johanna Mård, Hannah Mathew-Richards, Andrew McKenzie, Alfonso Mejia, Eduardo M. Mendiondo, Marjolein Mens, Shifteh Mobini, Guilherme S. Mohor, Viorica Nagavciuc, Thanh Ngo-Duc, Huynh T. T. Nguyen, Pham T. T. Nhi, Olga Petrucci, Nguyen H. Quan, Pere Quintana-Seguí, Saman Razavi, Elena Ridolfi, Jannik Riegel, Md S. Sadik, Nivedita Sairam, Elisa Savelli, Alexey Sazonov, Sanjib Sharma, Johanna Sörensen, Felipe A. A. Souza, Kerstin Stahl, Max Steinhausen, Michael Stoelzle, Wiwiana Szalińska, Qiuhong Tang, Fuqiang Tian, Tamara Tokarczyk, Carolina Tovar, Thi V. T. Tran, Marjolein H. J. van Huijgevoort, Michelle T. H. van Vliet, Sergiy Vorogushyn, Thorsten Wagener, Yueling Wang, Doris E. Wendt, Elliot Wickham, Long Yang, Mauricio Zambrano-Bigiarini, and Philip J. Ward. 2023. Panta rhei benchmark dataset: socio-hydrological data of paired events of floods and droughts. *Earth System Science Data*, 15(5):2009–2023.

Ni Li. 2023. Wikimpacts: Mining Wikipedia for climate impact information using machine learning. Master's thesis, KU Leuven.

Nominatim contributors. 2014. Place Ranking in Nominatim. https://nominatim.org/release-docs/latest/customize/Ranking/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya,

Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenStreetMap contributors. 2017a. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

OpenStreetMap contributors. 2017b. TMC/Location Code List/Location Types. https://wiki.openstreetmap.org/wiki/TMC/Location_Code_List/Location_Types.

Vikrant Panwar and Subir Sen. 2020. Disaster damage records of em-dat and desinventar: a systematic comparison. *Economics of disasters and climate change*, 4(2):295–317.

Dominik Paprotny, Pawel Terefenko, and Jakub Śledziowski. 2023. An improved database of flood impacts in europe, 1870–2020: Hanze v2.1. *Earth System Science Data Discussions*, 2023:1–37.

Jan Sodoge, Christian Kuhlicke, and Mariana Madruga de Brito. 2023. Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning. *Weather and Climate Extremes*, 41:100574.

Elisabeth Tschumi and Jakob Zscheischler. 2020. Countrywide climate features during recorded climate-related disasters. *Climatic change*, 158(3-4):593–609.

UNISDR. n.d. DesInventar: United Nations office for disaster risk reduction. Retrieved in May 2024 from https://www.desinventar.net.

Beichen Zhang, Frank Schilder, Kelly Helm Smith, Michael J. Hayes, Sherri Harms, and Tsegaye Tadesse. 2021. TweetDrought: A deep-learning drought impacts recognizer based on twitter data. In *Tackling Climate Change with Machine Learning Workshop at the Thirty-eighth International Conference on Machine Learning*. ICML.

A Keywords for Document Selection

Category	Keywords
Drought	drought, droughts, dryness, dry spell, dry spells, rain scarcity, rain scarcities, rainfall deficit, rainfall deficits, water stress, water shortage, water shortages, water insecurity, water insecurities, limited water availability, limited water availabilities, scarce water resources, groundwater depletion, groundwater depletions, reservoir depletion, reservoir depletions
Extreme Temperature	heatwave, heatwaves, heat wave, heat waves, extreme heat, hot weather, high temperature, high temperatures cold wave, cold waves, coldwave, coldwaves, cold snap, cold spell, arctic snap, low temperature, low temperatures, extreme cold, cold weather
Flood	floodwater, floodwaters, flood, floods, inundation, inundations, storm surge, storm surges, storm tide, storm tides
Wildfire	wildfire, forest fire, bushfire, wildland fire, rural fire, desert fire, grass fire, hill fire, peat fire, prairie fire, vegetation fire, veld fire
Storm	windstorm, windstorms, storm, storms, cyclone, cyclones, typhoon, typhoons, hurricane, hurricanes, blizzard, strong winds, low pressure, gale, gales, wind gust, wind gusts, tornado, tornadoes, wind, winds, lighting, lightings, thunderstorm, thunderstorms, hail, hails extreme rain, extreme rains, heavy rain, heavy rains, hard rain, hard rains, torrential rain, torrential rains, extreme precipitation, extreme precipitations, heavy precipitation, heavy precipitations, torrential precipitation, torrential precipitations, cloudburst, cloudbursts

Table 3: Keywords for document selection by event category. The category Storm subsumes the more specific categories Tornado, Tropical Storm/Cyclone, and Extratropical Storm/Cyclone in the database schema.

B Selected LLM Prompts

```
prompt_main_event=f'''
    Based on the provided article {info_box} {whole_text},
    please extract information about the main event {event_name},
    and assign the details as follows:

    - "Main_Event": "identify the event category referring to
    "Flood; Extratropical Storm/Cyclone; Tropical Storm/Cyclone; Extreme
    Temperature; Drought; Wildfire; Tornado".
    Only one category should be assigned."

    - "Main_Event_Assessment_With_Annotation": "Include text from
    the original text that supports your findings on the Main_Event."
    please give the json format output of these two items above,
    and please make sure that your annotation text is explicitly
    from the original text provided.
'''

prompt_country = f'''
    Based on the provided article {info_box} {whole_text},
    identify all countries affected by {event_name},
    and assign the appropriate details:
```

```
    - "Country": "List all countries mentioned in the text as being
    affected by {event_name}."

    - "Country_With_Annotation": "For each location listed, include
    a snippet from the article that supports why you consider it
    affected by {event_name}. This annotation should help illustrate
    how you determined the country was impacted. This should directly
    quote the original text."

    Please give the json format output of these two items above,
    and please make sure that your annotation text is explicitly
    from the original text provided.
'''

prompt_time = f'''
    Based on the provided article {info_box} {whole_text},
    identify the time infomation {event_name} described,
    and assign the appropriate details:

    - "Start_Date": "The start date of the event. If the specific
    day or month is not known, include at least the year if it's
    available. If no time information is available, enter 'NULL'.
    If the exact date is not clear (e.g., "summer of 2021", "June
    2020"), please retain the text as mentioned."

    - "End_Date": "The end date of the event. If the specific day or
    month is not known, include at least the year if it's available.
    If no time information is available, enter 'NULL'. If the exact
    date is not clear (e.g., "summer of 2021", "June 2020"), please
    retain the text as mentioned."

    - "Time_With_Annotation": "Include text from the original text
    that supports your findings on the start date and end date.
    This should directly quote the original text."

    Please give the json format output of these three items above,
    and please make sure that your annotation text is explicitly
    from the original text provided.
'''

prompt_death_per_country = f'''
    Based on the provided article, which includes the information
    box {info_box} and the full text {whole_text}, first extract
    and summarize the total number of deaths associated with
    {event_name}, along with supporting annotations from the article.
    Organize this information in JSON format as follows:

    - "Total_Summary_Death":{{
    - "Total_Deaths": "The total number of people who died in
    {event_name}, both directly and indirectly.
    Use the exact number if mentioned, or retain the text or range as
    provided for vague numbers (e.g., 'hundreds of,' '500 families,'
```

```
    'thousands of,' '300-500 people'). If the information is missing,
    assign 'NULL'."
    - "Total_Death_Annotation": "Provide excerpts from the article
    that directly support your findings on the total number of
    deaths. This should directly quote the original text."
    }}

    If the "Total_Deaths" is not "NULL" or "0", then, delve deeper to
    provide a detailed breakdown of these deaths by country.
    The first instance in the "Specific_Instance_Per_Country_Death"
    section for each country provides a summary of the total deaths
    within that country and the "Location_Death" is the country name,
    followed by a breakdown into specific cities, towns, or regions
    where possible. Organize this information in JSON format as follows:

    - "Specific_Instance_Per_Country_Death":[{{
    - "Country": "Name of the country."
    - "Location_Death": "The specific place within the country where
    the deaths occurred, including towns, cities, or regions."
    - "Start_Date_Death": "The start date when the deaths occurred,
    if mentioned."
    - "End_Date_Death":"The end date when the deaths occurred, if
    mentioned."
    - "Num_Death": "The number of people who died in this specific
    location or incident related to {Event_Name}. Use the exact
    number if mentioned, or retain the text or range as provided for
    vague numbers (e.g., 'hundreds of,' '500 families,' 'thousands
    of,' '300-500 people'). If the information is missing, assign
    'NULL'."
    - "Death_with_annotation": "Excerpts from the article that
    support your findings on the location, time, number of deaths.
    This should directly quote the original text."
    }}]

    Ensure to capture all instances of death mentioned in the
    article, including direct and indirect causes.
'''
prompt_total_per_country = f'''
    Based on the provided article, which includes the information
    box {info_box} and the full text {whole_text} related to
    {Event_Name}, first extract and summarize detailed information
    about the total economic loss or damage caused by {Event_Name},
    focusing specifically on the economic impact in the mentioned
    regions. The information should be organized in JSON format
    as follows:

    - "Total_Summary_Damage": {{
    - "Total_Damage": "Specify the economic loss or damage reported.
    If this information is not mentioned, assign 'NULL'."
    - "Total_Damage_Units": "Indicate the currency of the reported damage
    (e.g., USD, EUR). If the currency is not specified, assign 'NULL'."
    - "Total_Damage_Inflation_Adjusted": "State 'Yes' if the reported
```

```
    damage amount has been adjusted for inflation; otherwise, indicate
    'No'. If this aspect is not mentioned, provide your best judgment
    based on the context."
    - "Total_Damage_Inflation_Adjusted_Year": "Mention the year used for
    inflation adjustment, if applicable. If the amount is not adjusted
    for inflation or this detail is not provided, assign 'NULL'."
    - "Economic_Impact_with_annotation": "Directly quote portions of
    the text that substantiate your findings on the total economic loss
    or damage. This should directly quote the original text."
    }}

    If the "Total_Damage" is not "NULL" or "0", then, delve deeper
    to provide a detailed breakdown of economic damages by country.
    For the first instance in the
    "Specific_Instance_Per_Country_Economic_Damage" section for each
    country, provide a summary of the total economic damage within that country
    and the "Location_Damage" is the country name,
    followed by a breakdown into specific cities, towns, or regions
    where possible. Organize this information in JSON format as follows:
    - "Specific_Instance_Per_Country_Damage":[ {{
    - "Country": "Name of the country.",
    - "Location_Damage": "The specific place within the country where the
    economic impact occurred, including towns, cities, or regions."
    - "Damage": "The amount of economic damage."
    - "Damage_Units": "The currency of the economic damage, like USD, EUR.
    If not specified, assign 'NULL'."
    - "Damage_Inflation_Adjusted": "Indicate 'Yes' if the damage amount
    has been adjusted for inflation; otherwise, 'No'."
    - "Damage_Inflation_Adjusted_Year": "The year of inflation adjustment,
    if applicable. If not adjusted or not applicable, assign 'NULL'."
    - "Damage_Assessment_with_annotation": "Include text from the original
    article that supports your findings on the economic impact amount and
    details for each specific instance. This should directly quote the
    original text."
    }}]

    Ensure to capture all instances of economic loss or damage mentioned
    in the article, including direct and indirect causes, and organize
    them in the JSON format output.
'''
```

C Event Distributions in the Benchmark Database

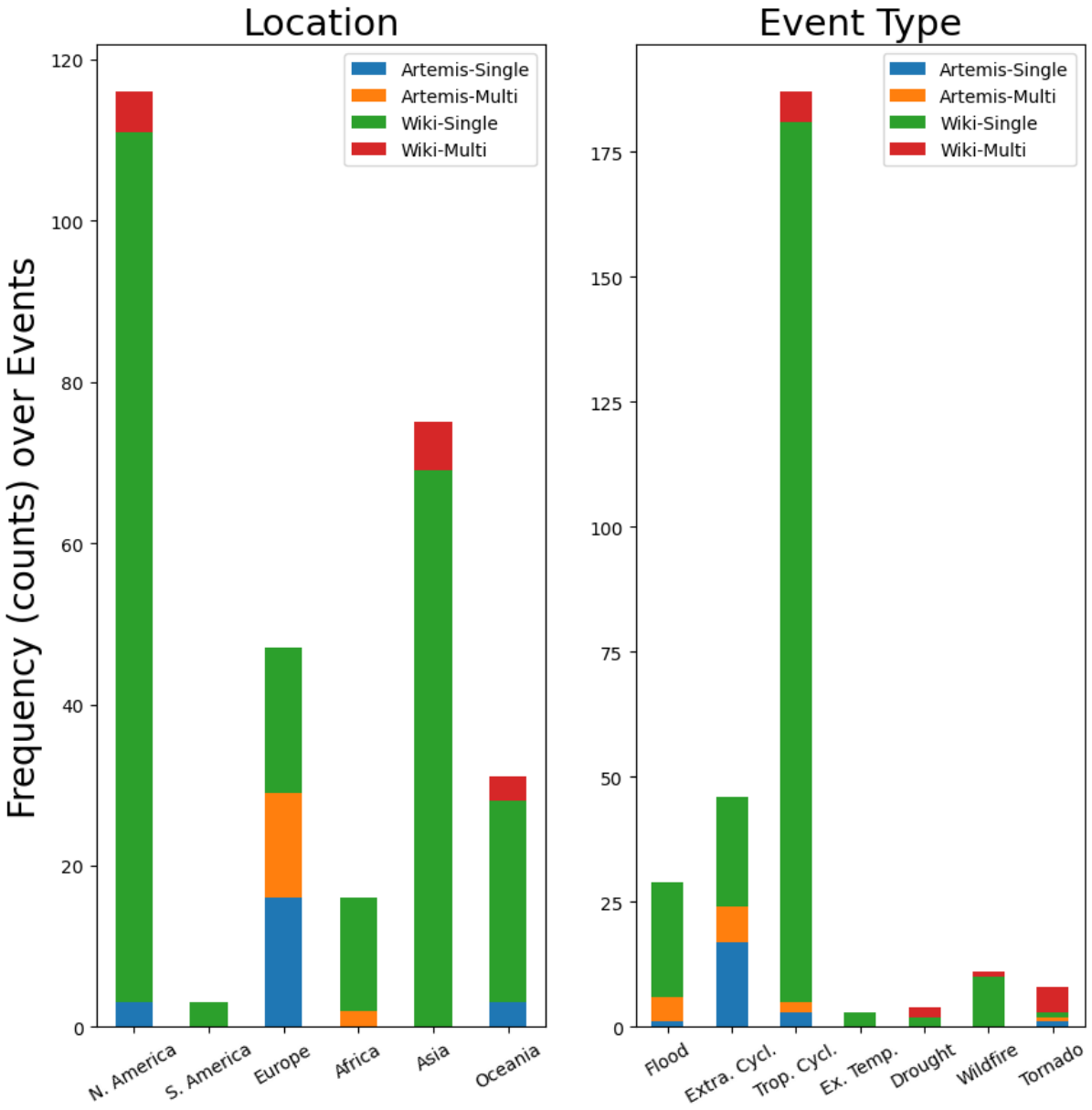

Figure 3: The left panel displays the co-distribution of event location in the benchmark database, categorized by the continent or large geographical region, with entry article source type, and frequency denoted by counts over the number (289) of database events. The right panel displays the same co-distribution, but for event category rather than location. *Extra. Cycl.* refers to the Extratropical Storm/Cyclone category, *Trop. Cycl.* to Tropical Storm/Cyclone, and *Ex. Temp.* to Extreme Temperature.

D Development Set Results

	GPT-4			Mistral (7B)			Mixtral (8x7B)		
Category	**Tot**	**Wik**	**Art**	**Tot**	**Wik**	**Art**	**Tot**	**Wik**	**Art**
Event Category	0.080	0.095	0.000	0.070	0.071	0.062	0.080	0.071	0.125
Location	0.335	0.310	0.466	0.466	0.415	0.730	0.479	0.454	0.609
Start-Year	0.020	0.012	0.063	0.740	0.809	0.375	0.130	0.095	0.312
Start-Month	0.049	0.047	0.063	0.750	0.821	0.375	0.150	0.120	0.312
Start-Day	0.103	0.058	0.339	0.753	0.822	0.400	0.208	0.160	0.455
End-Year	0.030	0.024	0.063	0.750	0.810	0.437	0.190	0.143	0.437
End-Month	0.058	0.048	0.112	0.760	0.821	0.437	0.205	0.160	0.442
End-Day	0.125	0.073	0.393	0.764	0.822	0.460	0.288	0.230	0.589
Deaths–Min	0.064	0.041	0.188	0.261	0.263	0.250	0.239	0.237	0.250
Deaths-Max	0.061	0.037	0.188	0.267	0.272	0.250	0.236	0.233	0.250
Damage-Min	0.191	0.110	0.617	0.490	0.526	0.304	0.334	0.267	0.687
Damage-Max	0.187	0.110	0.592	0.480	0.518	0.280	0.334	0.267	0.687
Damage-Cur	0.300	0.214	0.750	0.410	0.345	0.750	0.380	0.298	0.812
Total Event	0.115	0.084	0.274	0.497	0.523	0.364	0.232	0.195	0.426

Table 4: Results on the development set with three different LLMs: GPT-4, Mixtral, Mistral. Average difference over all events (Tot) and separately for Wikipedia (Wik) and Artemis (Art) articles.